

「2022년 고용노동데이터 활용 아이디어 공모전」

아이디어 기획서

아이디어명	머신러닝을 활용한 중소기업 채용정보 비교시스템 구축
아이디어 내용	
목 차	
0. 요약	2
1. 현황분석 및 아이디어 제안	4
1) 현황 및 문제상황	4
2) 아이디어 제안	6
(1) 사용한 데이터의 변수	6
(2) 아이디어 제안 방향	8
2. 아이디어 구현을 위한 모델링	10
1) 데이터의 수집 및 통합	10
2) 데이터 전처리	12
(1) A: 고용노동부 통합 데이터 전처리	13
(2) B: 중소기업 데이터 전처리	14
(3) 최종 데이터셋의 형성	15
3) 데이터 라벨링	15
4) 모델링	15
3. 아이디어 구현	23
1) 적용 방법	23
(1-1) 채용정보 비교시스템 구축	23
(1-2) 실제 데이터 적용	24
(2) 기준별 가중치를 이용한 조건부 검색 생성	25
(3) 워드 클라우드 활용한 기업 리뷰	26
(4) 채용 유형별 맞춤 정보 제공	27
2) 한계점	29

(1) 모델링의 한계	29
(2) 데이터 수집의 한계	29
(3) 홍보의 필요성	29
3) 기대효과 및 의의	30
(1) 구직자 측면	30
(2) 사회적 측면	31
4. 참고문헌	32

표 목 차

<표 2.3.1> 데이터 라벨링 평가지표, 활용변수, 등급분류 목록	16
<표 2.3.2> 데이터 라벨링 평가지표, Scree Plot, K-Means Clustering ..	17
<표 2.3.4> K-Means Elbow Diagram, Scatterplot	18
<표 2.4.1> 데이터 모델링 예측 지표, 독립 변수 목록	20
<표 2.4.2> train and Test Data	20
<표 2.4.3> 데이터 모델링 예측 지표, Accuracy Score(모델 평가지표) ..	22
<표 3.1.1> (주)더존시스템의 변수별 특성	24
<표 3.1.2> 비교시스템을 적용한 (주)더존시스템의 결과	25

그림목차

[그림 1.1] 워크넷에서 볼 수 있는 모 기업의 재무제표	5
[그림 2.1.1] 완성된 A 데이터셋의 일부	12
[그림 2.1.2] 완성된 B 데이터셋의 일부	12
[그림 2.2.1] KNN Imputer로 채워진 사망자 및 재해자수	13
[그림 2.2.2] 최종 데이터셋의 일부	15
[그림 2.3.3] PCA Scree Plot	18
[그림 2.3.5] 기업들의 종합점수 분포 (대략적인 정규분포 형태)	19
[그림 2.4.4] 기업들의 종합점수 지표 RMSE 그래프	22
[그림 3.1.3] 워크넷 시스템에 적용된 비교시스템 UI	25
[그림 3.1.4] 워크넷 시스템에 적용된 조건부 검색 UI	26
[그림 3.1.5] (주)빌컴의 최근 1년 간 리뷰에 대한 워드클라우드	27
[그림 3.1.6] 네이버 데이터랩의 '근로자건강센터', '외국인노동자지원센터'에 대한 3개월간의 빈도 통계	28
[그림 3.1.7] 워크넷 시스템에 적용된 채용 유형별 맞춤형 정보 UI	28
[그림 3.2.1] 2021년 3월 국내 구인, 구직 애플리케이션 및 사이트 이용 현황, 인크로스 ..	30

0. 요약

최근 중소기업의 인력난 문제가 점점 심해지고 있다. 그 원인 중 하나는 중소기업에 대한 부정적인 인식이다. 이러한 문제는 대기업이나 공공기관에 비해 부족한 중소기업의 정보 접근성으로부터 기인하였다. 설령 정보가 있더라도 대부분은 특정 기준에 대해 절대적인 수치만을 제공할 뿐 기업 간 비교와 같은, 구직자가 실제로 원하는 정보를 제공하고 있는지에 대해 의문이 들었다. 이와 관련하여 고용노동부가 운영하는 워크넷, HRD-Net에 채용정보 비교시스템을 도입하여 현재 발생하고 있는 ‘일자리 미스매치 현상’을 해결할 수 있도록 개선할 것이다.

보고서는 크게 현황분석 및 아이디어 제안, 아이디어 구현을 위한 모델링, 아이디어 적용, 총 3부분으로 구성하였다.

‘현황분석 및 아이디어 제안’에서는 현재 시스템의 현황을 분석하고 이를 개선하기 위해 아이디어를 간략히 설명하였다.

‘아이디어 구현을 위한 모델링’에서는 머신러닝 기법을 활용해, 구직자들의 유형을 고려한 여러 기준에 따라 기업을 직관적으로 평가할 수 있는 모델을 구현하였다.

‘아이디어 적용’에서는 모델링의 결과를 바탕으로 개선된 시스템의 구체적인 구현 방식에 대해 자세히 설명한다.

1. 현황분석 및 아이디어 제안

1) 현황 및 문제상황

대한민국에서 ‘중소기업’은 유독 부정적인 의미를 자주 내포한다. 낮은 임금, 열악한 처우, 불안한 고용 안정성 등을 자연스럽게 떠올리게 된다. 여러 커뮤니티만 보더라도 중소기업에 대한 부정적인 의견을 많이 볼 수 있다. 물론 이것이 단순히 ‘입소문’에 불과한 것은 아니다. 실제로 대기업이나 공공기관과 비교해볼 때, 여러 측면에서 차이가 나는 것이 사실이다.

그러나, 사회에 만연해있는 ‘중소기업’이란 단어의 선입견은 너무 과장되고, 지나치게 부정적이다. 우수한 청년들이 걸맞은 대우를 받으며 능력을 펼치는데 충분한 중소기업 역시 정말 많다. 하지만 한번 뇌리에 박힌 고정관념을 바꾸기란 쉽지 않을 것이다. 이러한 현상이 지속됨에 따라 관련된 문제들이 과생되고 있다.

중소기업의 인력난은 고용 문제에 있어서 대표적인 고질병이다. 한 기사에 따르면, 300인 미만 중소기업의 인력 부족률은 500인 이상 대기업의 7~8배에 달했고, 중견

기업과도 2배 이상의 차이가 난다고 한다. 부정적인 선입견으로 쉽사리 중소기업에 지원하지 않는 현상으로, 구직자와 채용자 사이의 ‘미스매치’가 일어나는 것이다.

이를 해결할 방법을 심도 있게 고민해보았다. 사회에 만연해있는 중소기업의 부정적인 편견을 바꾸기란 쉽지 않다. 하지만 구직자들이 중소기업을 선택하는 과정에 있어, 자세하고 정확한 정보를 제공한다면 그들의 인식 역시 달라질 것이다. 다른 기업과 비교했을 때 임금, 복지, 기업 평가 등의 수준이 얼마나 차이가 있는지 객관적으로 제시하는 것은 중소기업에 대한 구직자의 접근성을 높인다. 그리고 이것을 고용노동부의 워크넷이나 모바일 어플인 HRD-Net의 채용정보 형식을 개선하는 방식으로 구현하려고 한다.

<그림 1.1>은 현재 워크넷에 접속하면 볼 수 있는 채용광고 중, 모 기업의 재무제표를 보여준다. 이외에도 현재 워크넷에는 채용과 기업에 대한 여러 정보가 있다. 하지만 아쉬운 점이 있다면 ‘절대적인’ 수치가 대부분이라는 것이다. 수치를 보고 기업의 전반적인 수준을 파악할 수 있지만, 유사한 다른 기업과 비교했을 때 얼마나 차이가 있는지 ‘상대적인’ 수치는 제공되어 있지 않다.

예를 들어 재무에 대한 지식이 없는 사람은 <그림 1.1>의 재무제표를 본 후 인터넷에서 재무제표를 평가하는 방법을 다시 찾아봐야 한다. 하지만 다른 기업과 비교했을 때 재무제표의 수준이 어떤지 보여준다면 이러한 번거로움을 줄일 수 있다. 이렇듯 우리의 시스템은 중소기업에 대한 구직자의 접근성 확대를 위해 정보를 효율적으로 제시하는 방식을 중점적으로 고심하였다.

더불어 다양한 형태의 구직자에 주목하였다. 기업을 선정할 때 구직자의 형태에 따라 우선하는 기준이 다를 것이다. 예를 들어 장애인 구직자의 경우 임금의 수준보다도 근무환경의 안전성을 더 우선할 확률이 높다. 따라서 본 시스템을 구축할 때 다양한 기준을 적용하려고 한다. 단순히 임금, 복지뿐만이 아니라 근무안전성, 기업의 종합점수 등의 기준을 활용하여 구직자들이 개인 맞춤형으로 기업을 선택할 수 있게 할 것이다.

• 재무제표 [단위 : 천원]

구분	계정명	2019-12-31	2020-12-31	2021-12-31
대차대조표	자산총계	31,261,403	41,230,241	39,391,047
	부채총계	20,436,684	28,348,459	28,564,941
	자본총계	10,824,719	12,881,782	10,826,106
손익계산서	매출액	34,620,536	35,870,102	39,300,867
	영업이익	2,799,939	2,827,492	3,079,490
	당기순이익	2,002,607	2,057,063	2,206,769

+ 자세히보기

<그림 1.1. 워크넷에서 볼 수 있는 모 기업의 재무제표>

2) 아이디어 제안

(1) 사용한 데이터의 변수

기업을 평가하기 위해 총 13개의 변수를 자체적으로 만들어 비교하는 시스템을 구축하려고 한다. 13개의 변수에는 임금을 반영하는 업종규모대비임금액, 업종대비임금액, 전체대비임금액, 재무 정보를 반영하는 성장성, 수익성, 안정성, 활동성을 비롯하여 기업의 수준을 나타내는 중요한 변수들이 포함되어 있다. 변수들의 자세한 내용은 다음과 같다.

※ 임금과 관련된 기준의 경우 ‘평균 이하, 평균 이상’으로 분류하였다.

[1] 업종규모대비임금액

임금(월)을 업종 규모별 월평균임금으로 나눈 값으로, 동일 업종 및 동일 규모 기업 대비 해당 기업의 임금액 수준을 알 수 있는 변수이다.

[2] 업종대비임금액

임금(월)을 업종별 월평균임금으로 나눈 값으로, 동일 업종 대비 해당 기업의 임금액 수준을 알 수 있는 변수이다.

[3] 전체대비임금액

임금(월)을 전체 월평균임금의 평균으로 나눈 값으로, 전체 기업 대비 해당 기업의 임금액 수준을 알 수 있는 변수이다.

※ 성장성, 수익성, 안정성, 활동성의 경우 ‘상, 중, 하’로 분류한 후 각 수준을 의미하는 표현으로 명명했다.

(‘상중하’로 분류하는 경우 재무 상태를 온전히 반영하지 못할 것으로 추측했다)

[4] 성장성

재무제표의 순이익증가율 지표를 바탕으로 기업의 경영성과나 재무 상태가 전기 대비 당기에 얼마나 성장했는가를 평가하는 지표이다. 느슨한 성장, 점진적 성장, 비약적 성장 3개의 범주로 구성되어 있다.

[5] 수익성

재무제표의 총자산순이익을 지표를 바탕으로 일정기간 기업 활동의 최종적인 성과, 즉 손익의 상태를 측정하는 지표이다. 저수익, 중수익, 고수익 3개의 범주로 구성되어 있다.

[6] 안정성

재무제표의 부채비율 지표를 바탕으로 기업의 채무 변제 능력과 경기 변동 대처 능력을 판단할 수 있는 지표이다. 낮음, 보통, 높음 3개의 범주로 구성되어 있다.

[7] 활동성

재무제표의 총자본회전을 지표를 바탕으로 경영 활동을 위하여 취득한 특정 자산이 어느 정도 효율적으로 이용되었는가를 판단할 수 있는 지표이다. 저효율, 중효율, 고효율 3개의 범주로 구성되어 있다.

※ 근무 안전성, 수평적 조직문화, 위라벨, 자아실현의 경우 ‘상, 중, 하’로 분류하였다.

[8] 근무안전성

사망만인율과 천인율 지표를 바탕으로 근무의 안전성을 파악할 수 있는 지표이다. 상, 중, 하 3개의 범주로 구성되어 있다.

[9] 수평적 조직문화

채용기업의 수평적 조직문화에 대한 리뷰 점수를 통해 도출한 변수이며, 해당 기업이 얼마나 수평적인 조직문화를 가지는지 판단할 수 있는 지표이다. 상, 중, 하 3개의 범주로 구성되어 있다.

[10] 위라벨

채용기업의 위라벨에 대한 리뷰 점수를 통해 도출한 변수이며, 해당 기업의 업무 강도나 휴식 시간이 어느 정도인지 판단할 수 있는 지표이다. 상, 중, 하 3개의 범주로 구성되어 있다.

[11] 자아실현

채용기업의 자아실현에 대한 리뷰 점수를 통해 도출한 변수이며, 해당 기업에서 승진 기회 및 가능성을 판단할 수 있는 지표이다. 상, 중, 하 3개의 범주로 구성되어 있다.

[12] 기업종합평가지수

앞서 언급한 순이익증가율, 총자산순이익율, 부채비율, 총자본회전율 총 4가지 재무제표 지표를 모두 고려한 변수로써 해당 기업의 재무 상태를 종합적으로 파악할 수 있는 지표이다. 보통, 양호, 우수 3개의 범주로 구성되어 있다.

[13] 종합점수

위의 변수들을 모두 종합적으로 점수화하여 최종 점수를 도출하였다. ‘평균 이하, 평균 이상’처럼 2개의 범주로 나뉘는 데이터의 경우 각각 30점, 70점을 할당하였으며 3개의 범주로 나뉘는 데이터의 경우 각각 13점, 33점, 53점을 할당하였다. 이를 12로 나누어 100점 만점의 점수로 환산하였다. 즉 종합적으로 해당 기업을 판단할 수 있는 지표이다.

(2) 아이디어 제안 방향

[1] 채용정보 비교시스템 구축

상대적으로 정보량이 적은 중소기업의 구직자들을 위해, 구직자 유형을 고려한 13개의 지표를 활용하여 기업 간 ‘상대적으로’ 비교할 수 있는 방식으로 시스템을 개선할 것이다. 기존 시스템의 역시 다양한 정보가 있지만, 단순히 특정 수치만을 가지고 채용기업을 선택하는 것은 쉬운 일이 아니기에 더 직관적인 정보들이 필요하다. ‘연봉 3000만원’이라는 정보보다 ‘제조업 기업 연봉 평균보다 20% 많음’이라는 정보가 더 쉽게 이해될 것이다. 이를 위해 머신러닝을 활용하여 각 기업에 대한 13개 기준의 등급을 예측하는 모델을 구축하고 이를 파이프라인으로 연결하여 통합된 시스템을 만들 것이다.

[2] 기준별 가중치를 이용해 조건부 검색 생성

구직자의 유형별로 기업을 평가할 때 각자 선호하는 기준이 다를 것이다. 따라서 각 기준에 대한 기업의 수준을 보여줌과 동시에 구직자가 선호하는 기준에 따라 조건에 맞추어 기업을 검색할 수 있는 기능을 추가하였다. 예를 들어 각 기준에 대한 등급을 설정하면 이와 연관된 기업들의 목록을 보여주는 형식이다. 또한, 기준별 가중치를 적용하였다. 만약 ‘월평균임금’이 ‘근무안전성’보다 더 우선시된다면, ‘월평균임금’ 기준에 가중치를 적용하여 동일한 조건 아래 월평균임금이 높은 순서대로 출력되는 방식을 고안했다. 이는 구직자들을 위한 맞춤형 검색 기능이라 할 수 있다.

[3] 워드 클라우드를 활용한 기업 리뷰

워크넷 채용공고와 연결되어있는 ‘잡플래닛’의 기업 리뷰 점수는 상당히 주관적인 수치라고 할 수 있다. 어떤 사람은 ‘수평적 조직문화’에 대해 평균이라고 생각하고 3점을 줄 수 있지만, 또 다른 사람은 낮음이라고 생각한 후 3점을 줄 수 있다. 같은 3점이지만 서로 다른 뜻을 의미한다. 따라서 이러한 리뷰를 더 객관적으로 판단하기 위해서, 텍스트로 된 기업 리뷰를 워드클라우드로 표현하려고 한다. 워드클라우드란 텍스트의 단어의 빈도를 시각적으로 표현하는 기법이다. 이를 통해 기업 리뷰 점수와 함께 현직자들의 평가를 확인할 수 있는 지표를 구현할 것이다.

덧붙여, 현재 많은 구인, 구직 사이트에서 인공지능을 활용한 채용 추천 서비스를 많이 사용한다. 하지만 이러한 기능은, 구직자가 확인할 수 있는 기업이 다소 한정적이며, 자신의 관심사와 관련된 기업을 궁금해하고, 직접 찾아보고 싶은 구직자에게는 효율적이지 않을 것으로 판단했다. 따라서 그들에게 집약적인 정보를 제공함으로써 도움을 줄 수 있는 시스템을 개발하기 위해 고민하였다.

2. 아이디어 구현을 위한 모델링

※본 분석 및 모델링에는 Python을 사용했다.

앞에서 설명한 1.업종규모대비임금액 2.업종대비임금액 3.전체대비임금액 4.성장성 5.수익성 6.안정성 7.활동성 8.근무안전성 9.수평적 조직문화 10. 위라벨 11.자아실현 12.기업종합평가지수 13.종합점수, 총 13가지의 기준에 대한 등급을 머신러닝 기법을 활용하여 예측하는 모델을 구현했다.

1) 데이터의 수집 및 통합

모델링을 위해 수집된 데이터는 다음과 같다:

*데이터 출처는 참고문헌에 기재.

-
- (1) 산업_규모별_임금_및_근로시간.csv
 - (2) 한국산업안전보건공단_산업중분류별규모별사고사망자수.csv
 - (3) 한국산업안전보건공단_산업중분류별규모별질병사망자수.csv
 - (4) 한국산업안전보건공단_산업중분류별규모별사망만인율.csv
 - (5) 한국산업안전보건공단_산업중분류별규모별사고재해자수.csv
 - (6) 한국산업안전보건공단_산업중분류별규모별질병재해자수.csv
 - (7) 한국산업안전보건공단_산업중분류별규모별천인율.csv
 - (8) 2022년 고용노동부 선정 강소기업_(워크넷팀 제공)V2.csv
-

(1)의 “산업분류(기업의 업종)”와 “규모(직원수를 기반으로 한 기업의 규모)”, 두 변수를 바탕으로 (2)~(7) 데이터를 순차적으로 통합하여 “A”라는 하나의 데이터를 형성한다.

A 데이터는 (1)~(7)까지 모든 정보를 담고 있으며 변수는 기업들의 “업종”, “규모”, “전체임금총액(월급)”, “질병사망자”, “사고사망자”, “사망만인율”, “사고재해자”, “질병재해자”, “천인율”로 총 9가지다.

(8)의 경우 2022년 고용노동부의 워크넷 기관에서 선정한 16,655개의 우수 강소기업 명단이다. 이 데이터에는 해당 기업들의 “브랜드명”, “사업장명”, “사업자등록번호”, “대표자명”, “업종(중분류)”, “업종(소분류)”, “주소” 등 전반적인 기본 기업정보를 포함하는 변수들이 있다.

본 기획서에서는 분석의 편리를 위해 (8)의 16,655개의 기업 데이터를 전부 사용하지 않고 크기가 500인 표본을 엑셀의 RANDBETWEEN() 함수를 사용하여 무작위 추출하였다(이하 “B 데이터”라 칭함).

하단에 기재할 변수들은 기업에 대한 추가적인 여러 정보가 담긴, B 데이터에 새로 추가한 변수들이다:

※변수 생성 시 기본적으로 워크넷 기업정보 데이터를 참고하되, 결측 자료는 잡플래닛 혹은 잡코리아에 관련 데이터가 존재할 경우, 이를 크롤링하여 추가한다.

소재지: B의 “주소” 변수에서 시/도에 해당하는 앞부분만 추출

직원수: 잡코리아에서 해당 기업 정보 검색 → “사원수” 항목 참고

임금(월): 잡코리아에서 해당 기업 연봉 정보 검색 → 전체 평균 연봉 / 12

수평적_조직문화: 워크넷에서 해당 기업 검색 → 기업의 ‘사내문화’ 지표
1~5까지 Ranking이 되어 있으며 지표가 높을수록 조직문화가 수평적임.

워라밸: 워크넷에서 해당 기업 검색 → 기업의 ‘업무와 삶의 균형’ 지표
1~5까지 Ranking이 되어 있으며 지표가 높을수록 워라밸 수준이 좋음.

자아실현: 워크넷에서 해당 기업 검색 → 기업의 ‘승진기회및가능성’ 지표
<기업 회계 관련 지표> → 워크넷 홈페이지에 해당 기업 검색 후 [재무비율]
참고

순이익증가율: 기업의 성장성 지표

총자산순이익율(ROA): 기업의 수익성 지표

부채비율: 기업의 안정성지표

총자본회전율: 기업의 활동성 지표

위 과정까지 마무리하면 A와 B의 총 2개의 데이터가 완성된다.

A의 변수(9가지): 업종, 규모, 전체임금총액, 질병사망자, 사고사망자, 사망만인율, 사고재해자, 질병재해자, 천인율

B의 변수(18가지): 연번, 브랜드명, 사업장명, 사업자등록번호, 대표자명, 업종(중분류), 업종(소분류), 주소, 직원수, 임금(월), 소재지, 수평적_조직문화, 순이익증가율, 총자산순이익율, 부채비율, 총자본회전율, 워라밸, 자아실현

연도	분야	사업명	사업자(별 4대사업)	업종(중공업·업종(소부)·주조	직원수	임대(월)	조세지	수령액 조지분율	순이익(업종)	중상순이익율	부채비율	중상순이익률	회귀율	지표값
2022-1201	국유재산·토지개발사업	국유재산·토지개발사업	국유재산·토지개발사업	국유재산·토지개발사업	10	2500000000	국유재산·토지개발사업	3	75	2500000000	0.2	4	1	1
2022-00201	신항구	신항구	신항구	신항구	3	2537583	신항구	3	75	2537583	0.2	4	1	1
2022-12136	메인비즈기업	메인비즈기업	메인비즈기업	메인비즈기업	3	75	2537583	3	75	2537583	0.2	4	1	1
2022-14411	사회복지기업	사회복지기업	사회복지기업	사회복지기업	3	75	2537583	3	75	2537583	0.2	4	1	1
2022-05639	인도비즈기업	인도비즈기업	인도비즈기업	인도비즈기업	3	75	2537583	3	75	2537583	0.2	4	1	1
2022-11022	메인비즈기업	메인비즈기업	메인비즈기업	메인비즈기업	3	75	2537583	3	75	2537583	0.2	4	1	1
2022-13390	메인비즈기업	메인비즈기업	메인비즈기업	메인비즈기업	3	75	2537583	3	75	2537583	0.2	4	1	1
2022-02642	인도비즈기업	인도비즈기업	인도비즈기업	인도비즈기업	3	75	2537583	3	75	2537583	0.2	4	1	1
2022-16507	안전보건경영시스템 인증	안전보건경영시스템 인증	안전보건경영시스템 인증	안전보건경영시스템 인증	3	75	2537583	3	75	2537583	0.2	4	1	1
2022-03795	인도비즈기업	인도비즈기업	인도비즈기업	인도비즈기업	3	75	2537583	3	75	2537583	0.2	4	1	1
2022-00767	신항구	신항구	신항구	신항구	3	2537583	신항구	3	75	2537583	0.2	4	1	1
2022-00172	신항구	신항구	신항구	신항구	3	2537583	신항구	3	75	2537583	0.2	4	1	1
2022-00800	대한민국 원자력 우등기업	대한민국 원자력 우등기업	대한민국 원자력 우등기업	대한민국 원자력 우등기업	3	75	2537583	3	75	2537583	0.2	4	1	1
2022-16540	안전보건경영시스템 인증	안전보건경영시스템 인증	안전보건경영시스템 인증	안전보건경영시스템 인증	3	75	2537583	3	75	2537583	0.2	4	1	1
2022-14734	국유재산·토지개발사업	국유재산·토지개발사업	국유재산·토지개발사업	국유재산·토지개발사업	3	75	2537583	3	75	2537583	0.2	4	1	1
2022-05631	인도비즈기업	인도비즈기업	인도비즈기업	인도비즈기업	3	75	2537583	3	75	2537583	0.2	4	1	1
2022-07007	인도비즈기업	인도비즈기업	인도비즈기업	인도비즈기업	3	75	2537583	3	75	2537583	0.2	4	1	1
2022-14795	국유재산·토지개발사업	국유재산·토지개발사업	국유재산·토지개발사업	국유재산·토지개발사업	3	75	2537583	3	75	2537583	0.2	4	1	1
2022-02098	인도비즈기업	인도비즈기업	인도비즈기업	인도비즈기업	3	75	2537583	3	75	2537583	0.2	4	1	1
2022-03427	인도비즈기업	인도비즈기업	인도비즈기업	인도비즈기업	3	75	2537583	3	75	2537583	0.2	4	1	1
2022-00240	신항구	신항구	신항구	신항구	3	2537583	신항구	3	75	2537583	0.2	4	1	1
2022-00203	신항구	신항구	신항구	신항구	3	2537583	신항구	3	75	2537583	0.2	4	1	1
2022-05629	인도비즈기업	인도비즈기업	인도비즈기업	인도비즈기업	3	75	2537583	3	75	2537583	0.2	4	1	1
2022-15060	국유재산·토지개발사업	국유재산·토지개발사업	국유재산·토지개발사업	국유재산·토지개발사업	3	75	2537583	3	75	2537583	0.2	4	1	1
2022-00054	신항구	신항구	신항구	신항구	3	2537583	신항구	3	75	2537583	0.2	4	1	1
2022-14798	인도비즈기업	인도비즈기업	인도비즈기업	인도비즈기업	3	75	2537583	3	75	2537583	0.2	4	1	1
2022-03742	국유재산·토지개발사업	국유재산·토지개발사업	국유재산·토지개발사업	국유재산·토지개발사업	3	75	2537583	3	75	2537583	0.2	4	1	1
2022-00044	신항구	신항구	신항구	신항구	3	2537583	신항구	3	75	2537583	0.2	4	1	1
2022-00098	신항구	신항구	신항구	신항구	3	2537583	신항구	3	75	2537583	0.2	4	1	1
2022-08663	인도비즈기업	인도비즈기업	인도비즈기업	인도비즈기업	3	75	2537583	3	75	2537583	0.2	4	1	1
2022-00464	신항구	신항구	신항구	신항구	3	2537583	신항구	3	75	2537583	0.2	4	1	1

1.업종	1.규모	1.전체임금총액	2.질병사망자	2.사고사망자	2.사망만인율	3.사고재해자	3.질병재해자	3.천인율
B.광업(05~08)	1~4인	3803220	22	4	907.58	24	121	349.59
B.광업(05~08)	5~9인	3347485	15	0	359.4	18	87	162.895
B.광업(05~08)	10~29인	4360906	18	2	1108.515	52	196	523.115
B.광업(05~08)	30~99인	4761588	56	2	513.165	24	434	523.07
B.광업(05~08)	100~299인	NA	70	1	554.345	4	582	544.335
B.광업(05~08)	300인이상	NA	159	0	268.78	11	1783	381.86
C.제조업(10~34)	1~4인	2511923	58	42	4.645	5592	880	16.237
C.제조업(10~34)	5~9인	3200845	29	25	1.287	3748	664	9.78
C.제조업(10~34)	10~29인	3547949	85	46	2.37	6220	1239	8.612
C.제조업(10~34)	30~99인	3811756	70	39	3.112	4322	1145	6.162
C.제조업(10~34)	100~299인	4457941	22	17	0.6625	2041	951	4.403
C.제조업(10~34)	300인이상	5550422	64	15	2.432	2342	2565	3.919
D.전기, 가스, 증기 및 공기 조절 공급업(35)	1~4인	3915265	0	0	0.99	7	1	6.49
D.전기, 가스, 증기 및 공기 조절 공급업(35)	5~9인	5280389	0	0	0.18	4	1	1.2
D.전기, 가스, 증기 및 공기 조절 공급업(35)	10~29인	5401114	0	0	0.445	26	6	5.255
D.전기, 가스, 증기 및 공기 조절 공급업(35)	30~99인	5800501	1	0	0.84	35	6	1.96
D.전기, 가스, 증기 및 공기 조절 공급업(35)	100~299인	5254601	0	0	0.39	16	7	0.7
D.전기, 가스, 증기 및 공기 조절 공급업(35)	300인이상	5759473	3	0	0.35	14	7	0.48
E.수도, 하수 및 폐기물 처리, 원료 재생업(36~39)	1~4인	2758452	0 NA	NA		7 NA	NA	
E.수도, 하수 및 폐기물 처리, 원료 재생업(36~39)	5~9인	3189692	0 NA	NA		4 NA	NA	
E.수도, 하수 및 폐기물 처리, 원료 재생업(36~39)	10~29인	3918241	0 NA	NA		26 NA	NA	
E.수도, 하수 및 폐기물 처리, 원료 재생업(36~39)	30~99인	4122507	1 NA	NA		35 NA	NA	
E.수도, 하수 및 폐기물 처리, 원료 재생업(36~39)	100~299인	4267403	0 NA	NA		16 NA	NA	
E.수도, 하수 및 폐기물 처리, 원료 재생업(36~39)	300인이상	NA	3 NA	NA		14 NA	NA	

2) 데이터 전처리

(1) A: 고용노동부 통합 데이터 전처리

[1] 파생변수의 생성

먼저 “규모”를 범주형 파생변수로 재설정한다. 1~4인 → A, 5~9인 → B, 10~29인 → C, 30~99인 → D, 100~299인 → E, 300인 이상 → F로 “A/B/C/D/E/F”의 범주를 갖도록 ‘규모’를 설정한다.

[2] 결측치 처리

“질병사망자”, “사고사망자”, “사고재해자”, “질병재해자” 변수들의 결측치는 KNN Imputer 방법론을 활용하여 처리했다. KNN Imputer는 가까운 이웃의 수를 정하고 그 이웃들을 이용하여 결측치를 채우는 방식이다.

	1.업종	1.규모	1.전체임금총액	2.질병사망자	2.사고사망자	2.사망만인율	3.사고재해자	3.질병재해자	3.천인율	규모
0	B.광업(05~08)	1~4인	3805220.0	22.0	4.0	907.5800	24.0	121.0	349.590	A
1	B.광업(05~08)	5~9인	3347485.0	15.0	0.0	359.4000	18.0	87.0	162.895	B
2	B.광업(05~08)	10~29인	4360906.0	18.0	2.0	1108.5150	52.0	196.0	523.115	C
3	B.광업(05~08)	30~99인	4761588.0	56.0	2.0	513.1650	24.0	434.0	523.070	D
4	B.광업(05~08)	100~299인	NaN	70.0	1.0	554.3450	4.0	582.0	544.335	E
...
97	S.협회 및 단체, 수리 및 기타 개인 서비스업(94~96)	5~9인	2618072.0	2.0	1.0	0.9100	236.0	55.0	5.160	B
98	S.협회 및 단체, 수리 및 기타 개인 서비스업(94~96)	10~29인	3085493.0	1.0	1.0	0.4325	244.0	65.0	3.505	C
99	S.협회 및 단체, 수리 및 기타 개인 서비스업(94~96)	30~99인	3383215.0	1.0	0.0	0.2675	137.0	32.0	2.830	D
100	S.협회 및 단체, 수리 및 기타 개인 서비스업(94~96)	100~299인	4674394.0	4.0	1.0	0.3400	72.0	17.0	1.290	E
101	S.협회 및 단체, 수리 및 기타 개인 서비스업(94~96)	300인 이상	4306688.0	5.0	1.0	0.2730	30.0	17.0	1.900	F

<그림 2.2.1. KNN Imputer로 채워진 사망자 및 재해자수>

추가적으로 “사망만인율”과 “천인율” 결측치의 경우 각각 (질병사망자, 사고사망자), (사고재해자, 질병재해자)를 독립변수들로 하는 다중회귀 방법론을 사용하여 처리했다.

“전체임금총액”의 결측치 대체의 경우, 우선 업종별 평균 전체임금총액 값을 구하여 내림차순으로 정렬한다. 전체임금총액 결측 데이터들의 업종 위치를 정확히 파악한다. 그 후, 해당 업종의 전후에 있는 업종들의 평균값으로 결측치를 대체한다.

(2) B: 강소기업 데이터 전처리

[1] 결측치 처리

B 데이터에서 결측치를 처리해야 할 변수들은 다음과 같다:

“직원수”, “수평적_조직문화”, “순이익증가율”, “총자산이익율”, “부채비율”, “총자본회전율”, “위라벨”, “자아실현”, “임금(월)”.

→ 하단의 방식으로 결측치를 처리했다.

먼저 해당 변수의 데이터 중 결측치가 아닌 값의 업종별, 규모별 중앙값을 계산한다. 그 후, 앞서 구한 결측치들의 업종과 규모에 알맞게 중앙값으로 결측치들을 채운다.

*중앙값을 고려한 이유는 통계적 특성상 이상치의 영향을 덜 받기 때문에 평균보다 데이터의 특성을 더 잘 압축해서 표현하는 수치라고 할 수 있다.

[2] 파생 변수 생성

B에도 “규모”라는 범주형 파생변수를 형성한다. 기준은 전과 같이 1~4인 → A, 5~9인 → B, 10~29인 → C, 30~99인 → D, 100~299인 → E, 300인 이상 → F로 “A/B/C/D/E/F”의 범주를 갖도록 ‘규모’를 설정한다.

[3] 재무제표 데이터의 이상치 처리

특히 기업의 4가지 재무제표 데이터(순이익증가율, 총자산이익율, 부채비율, 총자본회전율)의 경우 이상치를 제거했다. 그 이유는 기업 간의 수치 차이가 크기 때문에 결과가 왜곡될 수 있기 때문이다.

여기서 이상치란 데이터를 오름차순으로 나열했을 때 $Q1 - 1.5 * IQR$ 보다 작거나 $Q3 + 1.5 * IQR$ 보다 큰 수치들을 말한다. ($Q1$ 은 25% 분위수, $Q3$ 75% 분위수, $IQR = Q3 - Q1$) IQR 방식을 활용하여 앞서 제시한 수치적 기준을 토대로 4가지 지표들의 이상치들을 전부 제거한다.

(3) 통합된 데이터셋 생성

전처리 과정이 끝났다면 최종적으로 A와 B 데이터의 공통 변수 “규모”와 “업종”을 기준으로 두 데이터를 하나로 통합한다.

최종 전체 데이터셋의 변수(19가지): 연번, 브랜드명, 사업장명, 사업자등록번호, 대표자명, 업종, 업종(소분류), 주소, 직원수, 임금(월), 소재지, 수평적_조직문화, 순이익증가율, 총자산순이익율, 부채비율, 총자본회전율, 워라벨, 자아실현, 월평균임금

	연번	브랜드 대명	사업장명	사업자등록 번호	대표 자명	업종	업종(소분 류)	주소	직원 수	임금(월)	소재 지	수평적 조직 문화	순이익 증가율	총자산 순이익 율	부채 비율	총자 본회 전율	위 라 벨	자아 실현	규모	월평균 임금	
	0	2022-00201	신형 기업	호성정과주 식회사	5048113350	김형 수	G.도매 및 소매업 (45~47)	상품 중개 업	대구 북구 대천로18길 34(대천동)	37.0	2537583.333	대구 광역시	3.0	-20.50	3.3	96.5	0.7	3.0	3.0	D	4474436.0
	1	2022-12136	메인 비즈 기업	황남빵	5050394362	최상 은	I.숙박 및 음식점업 (55~56)	음식점업	경상북도 경주시 태종 로 783	27.0	2500000.000	경상 북도	3.0	-75.10	2.7	205.5	0.5	2.0	2.0	C	2377333.0
	2	2022-05639	이노 비즈 기업	현빈개발 (주)	2148723087	권재 훈	S.협회 및 단체, 수리 및 기타 개인 서비스 업(94~96)	개인 및 가 정용품 수 리업	서울 서초구 논현로 79710호 (양재동, 윈도 스톤호피스빌딩)	21.0	2500000.000	서울 특별시	2.0	11.60	5.1	81.8	1.9	2.0	2.3	C	3085493.0
	3	2022-11022	메인 비즈 기업	합자회사 보 령환경	3138101700	문영 최대운	E.수도, 하수 및 폐기 물 처리, 원료 재생업 (36~39)	폐기물 수 집, 운반업	충남 보령시 신설3길 26-3(동대동)	29.0	2500000.000	충청 남도	3.0	74.00	8.6	61.1	0.6	3.0	2.0	C	3918241.0
	4	2022-13390	메인 비즈 기업	합자회사 건 우개발	5118112266	신윤 교	B.광업(05~08)	토사석 광 업	경북 문경시 산북면 운 달로 388건우개발	17.0	2500000.000	경상 북도	3.0	-75.75	3.7	51.0	0.6	2.8	2.7	C	4360906.0
	
	348	2022-14240	사회 적기 업	(유)열린사 회서비스센 터	2248142129	백영 화	Q.보건업 및 사회복지 서비스업(86~87)	비거주 복 지시설 운영업	강원 횡성군 횡성읍 어 사대로 41	64.0	2500000.000	전라 북도	2.7	58.40	23.5	111.1	2.8	2.8	2.8	D	3041902.0
	349	2022-10443	메인 비즈 기업	(유)성원엘 리베이터	4028186315	남미 현 양 해정	S.협회 및 단체, 수리 및 기타 개인 서비스 업(94~96)	산업용 기 계 및 장비 수리업	전라북도 전주시 덕진 구 신동안길 52	28.0	2615000.000	전라 북도	1.0	28.70	11.3	79.5	2.4	1.0	1.0	C	3085493.0
	350	2022-10796	메인 비즈 기업	(유)삼려환 경	4178116112	장정 근	E.수도, 하수 및 폐기 물 처리, 원료 재생업 (36~39)	폐기물 수 집, 운반업	전남 여수시 소라면 의 곡길 115	20.0	3024166.667	전라 남도	3.0	-39.10	7.8	31.7	0.6	2.0	2.0	C	3918241.0
	351	2022-14248	사회 적기 업	(유)나눔	2268138000	관순 환	Q.보건업 및 사회복지 서비스업(86~87)	비거주 복 지시설 운영업	강원도 강릉시 하평5길 9-0(포남동) 1층	63.0	2500000.000	강원 도	2.7	51.90	5.4	35.7	3.8	2.8	2.8	D	3041902.0
	352	2022-14456	사회 적기 업	(사) 대한생 물재육지도 자연협회	1108212075	이종 덕	R.예술, 스포츠 및 여 가관련 서비스업 (90~91)	스포츠 서 비스업	서울 서대문구 홍은중 영로 1381층 (홍은동)	33.0	2500000.000	서울 특별시	2.3	-4.90	-11.3	145.4	0.9	2.6	2.1	D	3502735.0

3) 데이터 라벨링 (기업 평가 지표 생성)

기준별로 기업들을 평가하기 위해서는 특정한 규칙을 정해 등급을 매기는 과정이 필요하다. 후에 모델링에 사용할 데이터는 독립변수와 종속변수가 존재해야 한다. 따라서 독립변수(하단 표의 활용변수)를 활용하여 각 기업에 대한 종속변수(하단 표의 평가 지표)의 등급을 생성해야 한다.

평가지표		활용 변수	등급분류
근무안전성		사망만인율, 천인율	상중하
월평균임금액 (평균 기준)	동일업종	임금(월), 월평균임금	평균이상, 평균이하
	동일규모 대비		
	동일업종 대비		
	전체업종 대비		
성장성		순이익증가율	비약적, 점진적, 느슨한
수익성		총자산순이익율	고수익, 중수익, 저수익
안정성		부채비율	높음, 보통, 낮음
활동성		총자본회전율	고효율, 중효율, 저효율
수평적_조직문화(상중하)		수평적_조직문화	상중하
위라벨(상중하)		위라벨	상중하
자아실현(상중하)		자아실현	상중하
기업종합평가지수 (재무제표 종합 평가)		순이익증가율, 총자산순이익율, 부채비율,	우수, 보통, 양호

	총자본회전율	
종합점수	모든 평가 지표 활용	100점 만점 中

<표 2.3.1. 데이터 라벨링 평가지표, 활용 변수, 등급분류 목록>

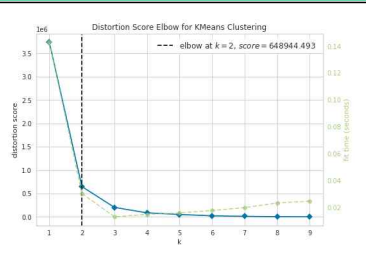
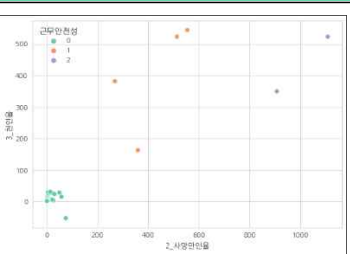
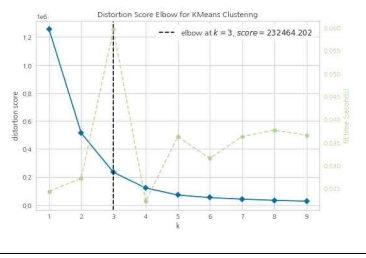
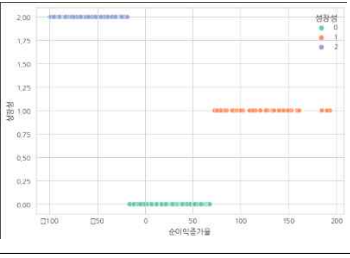
구현 과정에서 사용할 방법은 K-Means 기법이다.

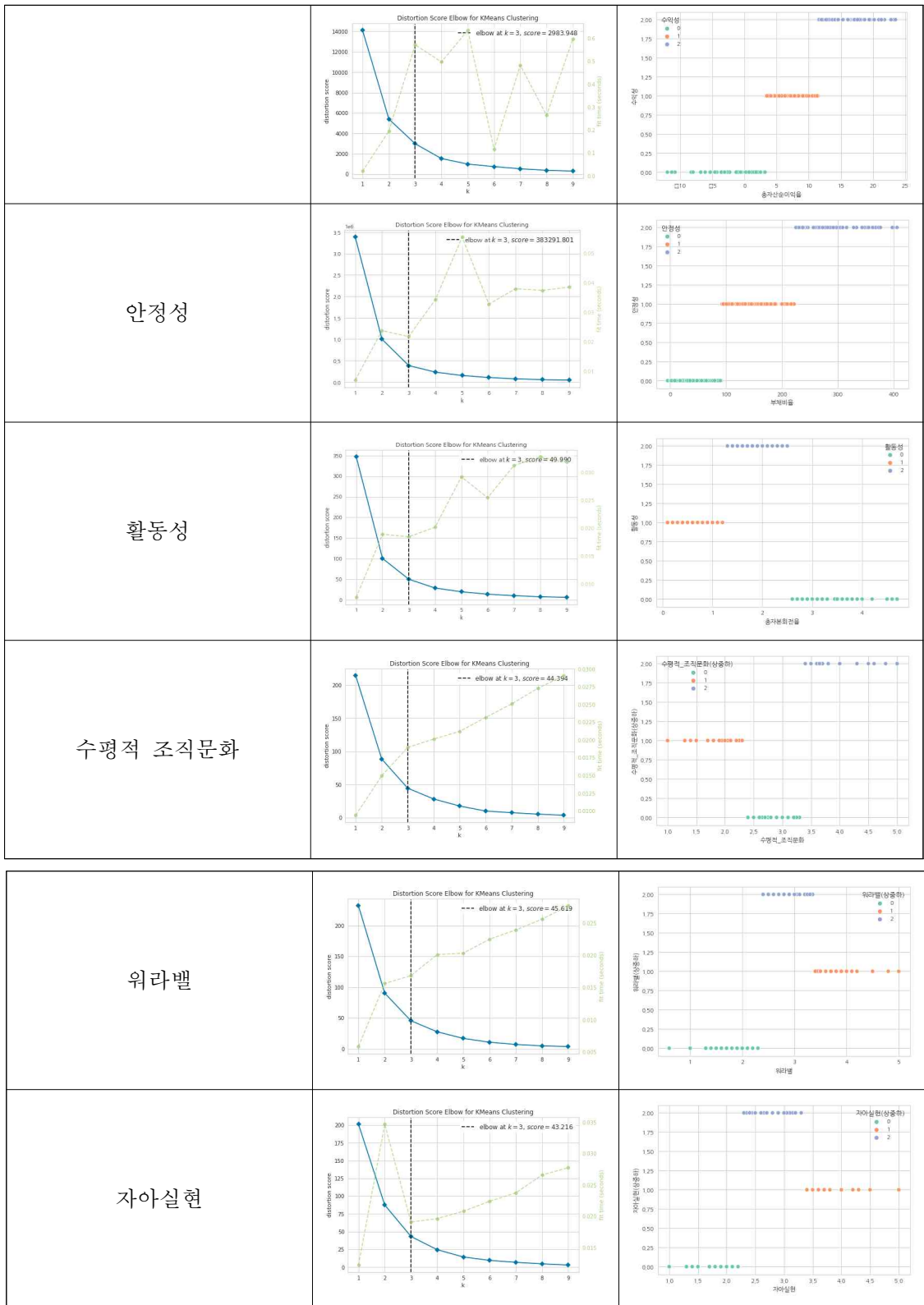
● K-Means?

머신러닝 비지도학습에 속하는 알고리즘으로, 주어진 변수를 사용하여 데이터를 K개의 군집(cluster)으로 묶는다. 각 군집의 평균을 활용하여 각 중심과 데이터들 사이의 거리를 계산하고 거리가 가깝게 위치하는 데이터를 비슷한 특성이 있는 데이터로 간주하여 군집화한다. 이때, K값은 곧 군집의 개수로, 여기서는 평가지표들의 범주 개수이다.

(1) 기업종합평가지수, 종합점수를 제외한 변수 라벨링

적절한 K값을 찾기 위해서는 “Elbow Method”가 활용된다. Elbow Method의 그래프를 통해 K값이 증가함에 따라 데이터 간 거리가 어떻게 변화하는지 파악할 수 있고, 여기서 그래프가 꺾이는 지점(표 2.3.2의 검정색 점선)이 적절한 K값이다. 각각 변수에 대한 그래프(Scree Plot)를 그려본 후 K=3으로 확정했다.

라벨링 지표명	그래프 (Scree Plot)	K-Means 클러스터링
근무안전성		
성장성		
수익성		



<표 2.3.2. 데이터 라벨링 평가지표, Scree Plot, K-Means Clustering>

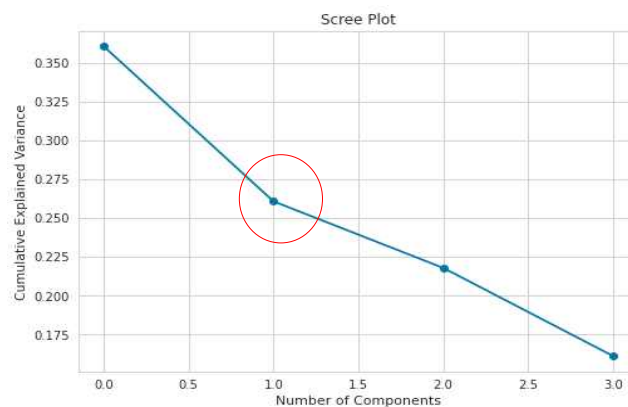
(2) 기업종합평가지수 라벨링

재무제표 4개의 지표(순이익증가율, 총자산순이익율, 부채비율, 총자본회전율)을 통합한 “기업종합평가지수” 변수를 만들었다. 그 후 4가지 지표를 모두 표준화(각기 다른 데이터의 평균과 표준편차를 표준 기수에 맞게 통일하는 작업)한다. 그 후 PCA(주성분분석)을 4가지 지표 데이터에 적용한다.

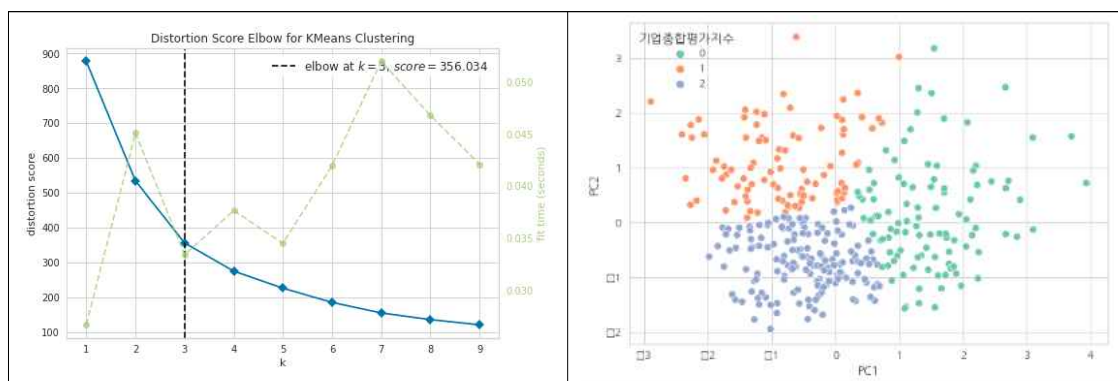
● PCA(주성분 분석)이란?

비지도학습 과정 중 하나로, 데이터에 변수가 너무 많으면 데이터를 좀 더 간결하게 표현할 수 있도록 차원을 축소하는 알고리즘이다. 차원 축소를 통해 데이터의 크기를 줄일 수 있고 비교적 시각화하기 쉽다는 장점도 있다.

본 재무제표 데이터에 PCA를 적용하면 Scree Plot의 Elbow Method(<그림 2.3.3 참고>)를 활용하여, 첫 번째 2개의 주성분(PC1, PC2)을 사용하는 것이 바람직함을 알 수 있다.



<그림 2.3.3. PCA Scree Plot>

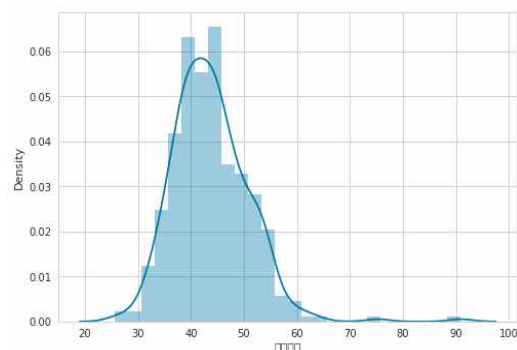


<표 2.3.4. K-Means Elbow Diagram, Scatterplot>

첫 2개의 주성분 데이터를 기반으로 K-Means 기법을 실행한 후 k=3임을 알 수 있고, 최종적으로 기업종합평가지수 데이터를 얻을 수 있다. 기업종합평가지수 산점도를 통해 3가지의 군집이 형성되었음을 볼 수 있고, 이를 이용해 기업종합평가지수 정도를 우수, 보통, 양호 수준으로 나누었다. 2개의 주성분의 PCA Score가 높을수록 기업종합평가지수 정도를 높다고 판단하였다.

(3) 종합점수 지표 라벨링

지금까지의 모든 변수를 종합하여 기업의 상대적 수준을 하나의 수치로 요약한 핵심 정보를 담고 있다. 앞서 제시한 표에서의 평가지표들을 점수화하여 모두 합산하면 된다. 점수화의 경우 범주가 3개인 변수는 범주별로(내림차순 기준) 53/33/13의 수치를 부여하고, 범주가 2개인 변수는 범주별로(내림차순 기준) 70/30의 수치를 부여했다.



<그림 2.3.5. 기업들의 종합점수 분포 (대략적인 정규분포 형태)>

4) 모델링

전처리와 라벨링이 완료된 데이터를 이용해 기준에 대한 등급을 예측하는 머신러닝 모델을 만들 것이다. 완성된 모델은 후에 새로운 데이터가 들어왔을 때 등급을 예측해주는 기능을 수행하게 된다.

<표 2.4.1>은 각각의 예측 지표에 따른 독립변수를 정리한 것이다.

예측 지표	독립 변수 (모델링에 활용되는 변수들)
근무안전성	업종, 규모, 사망만인율, 천인율
동일업종 동일규모 대비 임금액 수준	업종, 규모, 업종규모대비임금액
동일업종 대비 임금액 수준	업종, 규모, 업종대비임금액
전체업종 대비 임금액 수준	업종, 규모, 전체대비임금액
성장성	업종, 규모, 순이익증가율
수익성	업종, 규모, 총자산순이익율
안정성	업종, 규모, 부채비율

활동성	업종, 규모, 총자본회전율
수평적_조직문화(상중하)	업종, 규모, 수평적_조직문화
위라벨(상중하)	업종, 규모, 위라벨
자아실현(상중하)	업종, 규모, 자아실현
기업종합평가지수 (재무제표 종합 평가)	업종, 규모, 순이익증가율, 총자산순이익율, 부채비율, 총자본회전율
종합점수	모든 평가 지표 활용

<표 2.4.1. 데이터 모델링 예측 지표, 독립 변수 목록>

예측을 위한 머신러닝 알고리즘 설계에 있어 반드시 거쳐야 할 과정은 주어진 데이터를 “훈련용 데이터(Train Data)”와 “예측용 데이터(Test Data)”로 나누어야 한다는 점이다. 훈련용 데이터는 모델을 학습시키기 위해 사용되는 것이고, 예측용 데이터는 알고리즘의 성능을 평가하기 위해 준비된 데이터 일부를 미리 분리해야 한다. 본 기획에서는 500개라는 다소 적은 데이터의 특성상 Train 대 Test의 비율을 7:3으로 진행했다.

모델링 과정은 예측 지표별로 전부 유사하다. 예시로 “근무안전성” 지표를 예측하는 모델을 형성해보도록 하겠다.

- (1) 업종, 규모, 사망만인율, 천인율을 변수로 가지는 데이터를 이용한다.
- (2) 데이터를 전부 범주화하고 7:3의 비율로 각각 Train, Test 데이터로 나눈다.

	Train Data (70%)	Test Data (30%)
독립 변수	x: 업종, 규모, 사망만인율, 천인율	x: 업종, 규모, 사망만인율, 천인율
예측 변수	y: 근무안전성	y: 근무안전성

<표 2.4.2. Train and Test Data>

(3) 예측 변수인 근무안전성은 범주가 상중하로 분류되어 있다. 그런데 해당 지표는 범주별로 고르게 분포된 것이 아닌, 특정 범주에 데이터양이 더 많이 있는 불균형 상태일 가능성이 크다. 따라서 이런 불균형 문제를 해결하기 위해 SMOTE 기법을 활용한다.

● SMOTE?

SMOTE는 대표적인 데이터 불균형 해소 기법의 하나로, 낮은 비율을 차지하는 범주

의 데이터 수를 새롭게 생성하며 늘리면서 불균형을 해소하는 방법이다. SMOTE 기법을 활용하여 train data의 개수를 증가시켰다.

(4) Train Data를 기반으로 근무안전성 지표를 예측하는 모델을 형성한다. 이때 사용되는 머신러닝 모델은 Random Forest 기법이다. Train data를 Random Forest에 훈련시킨다. 모델의 성능을 높이기 위해 Grid search 기법을 활용하여 Random Forest 모델의 초모수를 조정하였다.

Random Forest 모델의 max_depth(최대 깊이), max_features(Feature의 개수), n_estimators(사용할 tree 개수)를 중심으로 최적의 초모수를 찾았다.

● Random Forest?

Random Forest란 지도 머신러닝 알고리즘의 하나로, 높은 정확도와 유연성으로 인해 가장 많이 사용되는 알고리즘 중 하나다. 특히 현재 데이터처럼 데이터의 범주가 광범위하고 변수들이 복잡한 시나리오에서는 모든 요소에 대한 의사 결정 트리를 생성하여 최종 결과 예측이 훨씬 정교하다.

● Grid search?

모델에게 가장 적합한 초모수 값을 찾기 위해, 여러 초모수들의 값을 변경해보며 최적의 성능을 보여주는 초모수를 보여준다.

● 초모수?

기계학습에서 학습의 기준이 되는 주요 변수로 주로 인간이 수작업으로 설정해야 한다.

(5) 모델 성능을 검증하기 위해 Test Data를 사용하여 정확도 측정(accuracy score)을 계산한다. 보통 Accuracy Score가 0.8~9 이상이면 Test Data 중 8~90% 이상이 적절한 예측에 성공했다는 뜻이므로 모델의 성능이 매우 좋다고 할 수 있다.

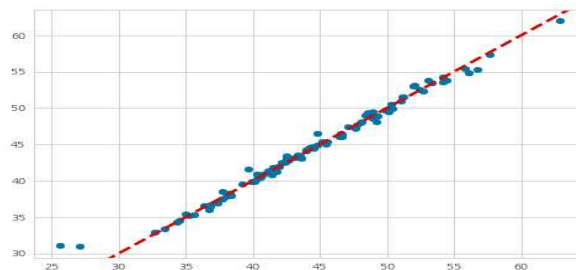
모델링 지표	Accuracy Score
근무안전성	1.0
동일업종 동일규모 대비 임금액 수준	1.0
동일업종 대비 임금액 수준	1.0
전체 업종 대비 임금액 수준	1.0
성장성	0.9811320754716981
수익성	0.9905660377358491
안정성	1.0
활동성	1.0

수평적 조직문화	1.0
위라벨	1.0
자아실현	1.0
기업종합평가지수	0.8679245283018868
종합점수	0.8249861066679886

<표 2.4.3. 데이터 모델링 예측 지표, Accuracy Score(모델 평가지표)>

모델링의 결과 대부분 지표에서 accuracy가 1에 수렴하는 성능을 보여주었다. 좋은 결과에도 불구하고, Random Forest 모델은 과대 적합에 취약하고, 훈련에 사용한 데이터의 개수가 적었기 때문에 바로 시스템에 적용하는 것은 성급할 수 있다. 따라서 실제 시스템에 적용하기 위해서는 최대한 많은 수의 표본을 확보하는 것이 중요할 것이다.

또한, 예측값이 등급분류가 아닌 점수로 출력되는 ‘종합점수’ 지표의 경우 RMSE(평균 제곱근 오차) 방식을 이용해 모델을 평가하였다.



<그림 2.4.4. 기업들의 종합점수 지표 RMSE 그래프>

그래프에서 볼 수 있듯이 매우 좋은 성능을 기록하였다. 이렇듯 총 13개의 지표에 대한 예측 모델을 구현해보았다.

3. 아이디어 적용

1) 적용방법

(1-1) 채용정보 비교시스템 구축

본 시스템은 중소기업 구직자들의 채용기업에 대한 손쉬운 접근을 위해 고안되었으며, 이를 활용해 워크넷과 HRD-Net에 게시된 기존 채용공고 시스템을 개선하려고 한다. 앞서 서술했던 문제 상황을 간략하게 요약한다면, 중소기업은 대기업이나 공공기관에 비해 상대적으로 기업에 대한 정보가 희소하다. 이러한 이유로 채용 과정에서 중소기업 구직자들은 정보의 부족으로 많은 어려움을 겪고 있다. 실제로, 고용노동부 모바일 어플인 ‘HRD-Net’의 ‘구인정보’, 고용노동부의 ‘워크넷’의 ‘채용정보’을 살펴보면 채용공고와 기업정보만을 보면서 기업을 온전히 이해하기란 쉽지 않다.

예를 들어, 기업정보에 게시된 재무에 대한 절대적 수치와 변화의 경우, 재무 지식이 있지 않은 이상, 기업의 재무수준이 어느 정도인지, 다른 기업과 어떤 차이가 있을지 단번에 확인할 수 없는 문제점이 있다. 위에서 예측 모델링을 진행한 여러 지표들 역시 이와 같은 문제점을 가지고 있다.

마찬가지로 워크넷의 ‘테마별 채용관’의 경우, 테마별로 우수한 기업들을 분류해 소개하고 있지만, 정작 그 안에 속하는 각개의 기업들의 수준을 일일이 비교, 평가할 수 없다. 따라서 각 기업을 평가할 수 있는 ‘상대적인’ 기준을 확인할 수 있는 ‘채용정보 비교시스템’을 고안하였다.

하단의 <그림 3.1.3> 과 같이 모바일이나 워크넷 홈페이지의 UI를 변경한 후, 앞서 개발했던 예측 모델들을 파이프라인으로 연결하며 비교시스템을 구축할 수 있다. 예측을 위한 머신러닝 기법을 적용함으로써, 홈페이지에 새로운 채용(기업)데이터가 추가되면 사전에 구축된 비교시스템에 의해, 기준에 대한 상대적 위치가 결정되는 원리이다. 채용정보 비교시스템을 이용하는 과정은 다음과 같다.

[1] (중소기업) 구직자가 채용정보를 확인하기 위해 ‘HRD-Net’이나 ‘워크넷’을 통해 ‘강소기업관(우수한 중소기업을 분류한 테마별 채용관)’의 A 기업 채용공고를 클릭한다.

[2] 시스템화되어 있는 채용정보 비교시스템에 의해 ‘채용정보 비교’ 항목에, 앞서 모델링한 13개의 기준에 대한 등급분류와 구체적인 상대적 위치가 출력된다.

[3] 이를 확인함으로써, 따로 기업에 대한 정보를 찾아보지 않아도 지원하고자 하는 기업의 상대적인 수준을 확인할 수 있다.

추가로, 지금까지 공공기관, 대기업에 비해 기업에 대한 정보가 부족한 중소기업을 대상으로 분석을 진행하였지만, 이러한 시스템을, 채용하려는 모든 기업에 적용해도 손쉽게 비교할 수 있다.

(1-2) 실제 데이터 적용

위의 채용정보 비교시스템을 실제 기업에 적용해보았다.

(주)더존시스템을 대상으로, <표 3.1.1>에 나와있는 각 독립변수의 값을 입력한 후 구현한 모델에 적용하였다.

사업장명	(주) 더존시스템
사망만인률	0.125
천인률	3.36
임금(월)	2653333.333
순이익증가율	-6.9
총자산순이익율	10
부채비율	106
총자본회전율	0.8
수평적_조직문화	1.3
워라밸	3
자아실현	2.5

<표 3.1.1. (주)더존시스템의 변수별 특성>

그 결과는 <표 3.1.2>와 같았다. 총 13가지 변수에 대한 등급과 상대적인 세부 점수 역시 확인할 수 있었다.

사업장명	(주) 더존시스템	
근무안전성	상	-
업종규모대비 임금액	평균 이하	동일 규모의 업종 대비 1,254,793원 낮습니다
업종대비 임금액	평균 이하	동일 업종 대비 1,548,571원 낮습니다
전체대비 임금액	평균 이하	전체 기업 대비 1,183,260원 낮습니다
성장성	점진적 성장	전체 기업 대비 12% 높습니다
수익성	중수익	전체 기업 대비 4.21% 높습니다.
안정성	보통	전체 기업 대비 10.6% 높습니다.
활동성	저효율	전체 기업 대비 0.68% 낮습니다.
수평적 조직문화	하	(5점 만점) 전체 기업 대비 1.4점 낮습니다.
위라벨	중	(5점 만점) 전체 기업 대비 0.25점 높습니다.
자아실현	중	(5점 만점) 전체 기업 대비 0.13점 높습니다.
기업종합평가지수	보통	-
종합점수	38	-

<표 3.1.2. 비교시스템을 적용한 (주)더존시스템의 결과>

<그림 3.1.3>은 채용 사이트의 시스템에 적용된 모습을 UI로 구현해본 것이다.



<그림 3.1.3. 워크넷 시스템에 적용된 비교시스템 UI>

(2) 기준별 가중치를 이용해 조건부 검색 생성

다음의 기능은 현재 ‘워크넷-채용정보’에서도 시행하고 있는 조건(필터링) 검색을 개선했다. 1번에서 제시한 13가지의 기준에 대한 등급을 각각 선택한 후 필터링하면 관련 기업들의 명단이 출력되는 방식이다. 예를 들어, 성장성은 ‘비약적 성장’, 위라벨은 ‘중’을 선택하게 되면 조건에 충족하는 기업들을 확인할 수 있을 것이다. 이는 구직하려는 기업에 대한 일정한 기준을 가지고 있는 구직자에게 적합한 기업을 찾는 시

간을 절약해줄 것이다.

이와 더불어 가중치가 반영된 조건 검색도 가능하다. 예를 들어 어떤 구직자가 ‘근무안전성’(상), ‘수익성’(고수익), ‘동일업종대비임금액’(평균이상), ‘안정성’(보통)을 기준으로 기업을 검색할 때, 4가지 기준 중 ‘근무안전성’에 더 높은 우선순위를 가지고 있다고 가정해본다.

이때 시스템은 ‘근무안전성’에 다른 3개의 기준보다 더 높은 가중치를 주게 되고, 4가지 조건 아래 ‘근무안전성’이 높은 순서대로 기업이 정렬된 후 출력되는 방식이다.

워크넷에서 ‘테마별 채용관’을 설립한 취지에서 유추할 수 있듯이, 구직자들의 유형별로 각자 우선시하는 기준들이 다르다. 그렇기에 기준별로 가중치를 적용한 조건부 검색은 개개인을 고려하는 ‘맞춤형’ 검색을 가능하게 한다. <그림 3.1.4.>는 이를 구현한 UI이다.

<그림 3.1.4. 워크넷 시스템에 적용된 조건부 검색 UI>

(3) 워드 클라우드를 활용한 기업 리뷰

현재 워크넷의 채용 사이트에서 채용기업에 대한 ‘기업 리뷰’를 볼 수 있다. ‘잡플래닛’ 사이트와 연동되어 ‘기업만족도’, ‘복지 및 급여’, ‘업무와 삶의 균형’ 등의 항목에 대한 리뷰 점수가 기재되어있다. 개인의 주관에 담긴 이러한 점수는 기업을 평가하기 상당히 까다롭다. 앞서 이 데이터를 이용해 등급을 분류하는 모델링을 수행하였지만, 원 데이터의 특성상 신뢰도가 떨어지는 것이 사실이다. ‘기업만족도’가 2.4점인 기업

이 2.5점인 기업보다 상대적으로 더 만족도가 높은 기업이라 단정하긴 힘들다. 잡플래닛과 연동하여, 현직자들이 각 기업에 관해 서술한 상세한 리뷰를 워드클라우드로 표현하면 위의 문제를 해결할 수 있다. 워드클라우드란 단어의 빈도를 시각적으로 나타내는 방법이다. 비록 각 항목에 대한 정확한 수치로 기업의 상대적인 위치를 파악하지는 못하겠지만, 리뷰에서 추출한 특정 단어의 빈도수를 확인하며 기업의 대략적인 분위기나 상황을 판단할 수 있을 것이다. <그림 3.1.5>는 모 기업에 대한 워드클라우드 예시이다.



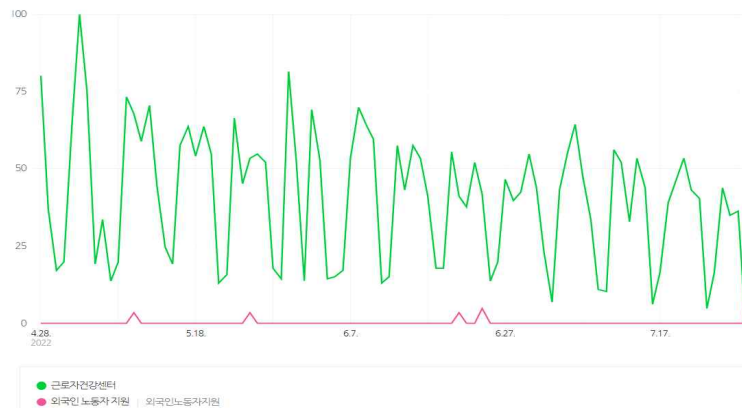
<그림 3.1.5. (주)빌컴의 최근 1년 간 리뷰에 대한 워드클라우드>

(4) 채용 유형별 맞춤 정보 제공

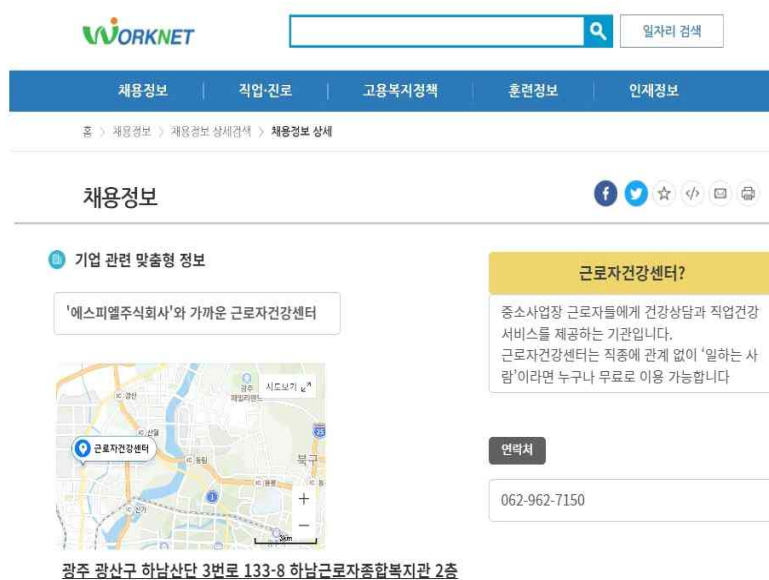
채용 화면에 비단, 기업에 관한 정보만 보여주는 것이 아니라, 이외의 유용한 정보 역시 제공한다면 채용 사이트를 이용한 구직자들의 정보 획득은 더욱 효율적일 것이다. 예를 들어 워크넷에 활성화되어 있는 ‘테마별 채용관’ 몇몇은 우수한 중소기업을 선정해 그 목록을 보여주고 있다. 이때 고용노동부에서 제공하는 근로자 건강센터 현황에 대한 공공데이터[고용노동부_근로자 건강센터 현황_20220401]를 연동시키면 미처 알지 못했던, 중소기업 근로자의 건강을 위한 정부의 지원수단이 존재함을 인지할 수 있을 것이다. 또한, 비슷하게 외국인 노동자에 대한 채용의 경우 고용노동부의 외국인 노동자 지원센터에 대한 공공데이터[고용노동부_외국인노동자 지원 센터_09/30/2021]를 연동한다면 외국인 노동자들이 이를 보고 해당 기관에 많은 도움을 받을 수 있을 것이다.

근로자 건강센터나 외국인 노동자 지원센터와 같이, 많은 사람이 서비스 대상에 포함되지만, 미처 알지 못해 이용하지 못한 기관이 많을 것이다. <그림 3.1.6.>은 지난 3개월간 ‘근로자 건강센터’와 ‘외국인 노동자 지원센터’의 검색어 빈도를 보여준다.

하루 평균 41번의 낮은 빈도를 보여주는 ‘근로자 건강센터’와 ‘외국인 노동자 지원센터’의 0에 수렴하는 빈도를 통해 많은 사람이 인지하지 못함을 확인할 수 있다. 이외에도 장애인, 고령자 등 해당 대상의 사람들도 미처 알지 못했던, 다양한 맞춤형 정보가 있을 것이다. 다음 <그림 3.1.7.>과 같이 이러한 정보를 채용공고와 연결하면 구직자들에게 유용한 정보를 손쉽게 제공할 수 있다.



<그림 3.1.6. 네이버 데이터랩의 ‘근로자건강센터’, ‘외국인노동자지원센터’에 대한 3개월 간의 빈도 통계>



<그림 3.1.7. 워크넷 시스템에 적용된 채용 유형별 맞춤형 정보 UI>

2) 한계점

(1) 모델링의 한계

앞서 채용정보 비교시스템의 프로토타입을 만들기 위해 500개의 데이터를 활용하였다. 7대 3의 비율로 Train, Test 데이터를 나누고, Train 데이터에 Smote 기법(낮은 비율로 존재하는 클래스의 데이터를 KNN 알고리즘을 활용하여 새롭게 생성하는 방법)을 활용하여 Train의 개수를 증가시켰다. Train 데이터를 통해 만들어진 예측 모델을 Test 데이터에 적용할 때 굉장히 좋은 성능을 보여주었다. 하지만 데이터의 개수가 적고, 과대적합(모델이 훈련 세트에서는 좋은 성능을 내지만 검증 세트에서는 낮은 성능을 내는 경우)의 가능성이 큰 ‘Random Forest’ 모델을 사용하였기에, 실제로 새로운 데이터를 적용할 때, 낮은 성능이 나올 확률이 높다. 그렇기에 이를 실제 시스템에 구현하기 위해서는 더 많은 수의 표본을 확보해야 할 것이다.

(2) 데이터 수집의 한계

기업 간의 정보를 비교하기 위해서는 각 기업의 정보를 가지는 데이터가 필요하다. 하지만 일부 규모가 작은 기업의 경우 재무제표에 대한 정보가 없거나, 기업 리뷰에 대한 점수를 확인할 수 없다. 그렇다면 기업 간의 비교 역시 제한적이다. 따라서 데이터를 보유하고 있는 유사 업종 기업의 정보를 제공하는 식의 대안을 통해 구직자에게 최대한 많은 정보를 전달할 수 있을 것이다.

(3) 홍보의 필요성

아무리 좋은 시스템을 구축하더라도, 사람들이 사용하지 않는다면 의미가 없을 것이다. [그림 3.2.1]과 같이 현재 수많은 채용 사이트가 있는 가운데, ‘워크넷’의 경우, 모바일 어플의 점유율은 10위권 밖에 있지만, 나름 Web 분야에서는 좋은 성과를 내고 있음을 확인할 수 있다. 하지만 이것에 안주하지 않고, 다양한 홍보 방법을 통해 워크넷이나 HRD-Net의 위와 같은 획기적인 시스템을 알린다면, 더 많은 사람에게 효율적인 정보를 제공할 수 있을 것이라 확신한다.

구인·구직순방문자수 TOP 10*					
(Unit: 만 명)					
APP			WEB		
1	알바몬	186.7	1	saramin.co.kr	208.5
2	사람인	133.4	2	work.go.kr	200.5
3	알바천국	133.0	3	jobkorea.co.kr	152.2
4	잡코리아	125.4	4	incruit.com	147.9
5	워크넷	85.9	5	albamon.com	46.9
6	잡플래닛	35.9	6	alba.co.kr	38.6
7	원티드	27.7	7	jobplanet.co.kr	32.1
8	인크루트	17.3	8	jobaba.net	11.7
9	자소설닷컴	16.6	9	catch.co.kr	11.2
10	벼룩시장구인구직	11.7	10	career.co.kr	10.4

<그림 3.2.1. 2021년 3월 국내 구인, 구직 애플리케이션 및 사이트 이용 현황, 인크로스>

3) 기대효과 및 의의

(1) 구직자 측면

위와 같은 시스템 개발은 구직자들의 정보 접근성을 확대한다. 구직 과정에서 가장 중요한 것 중 하나는 자신이 지원하려는 회사에 대해 아는 것이다. 그 회사가 어떤 회사인지, 어떻게 평가되는지는 구직자의 지원 의사결정에 크게 영향을 끼친다. 하지만 고용노동부의 워크넷을 비롯하여 많은 구인, 구직 사이트들은 기업에 대한 ‘절대적인’ 정보를 중점적으로 보여주고 있다. 물론 기업 리뷰가 몇 점인지, 매출액이 얼마인지 아는 것도 중요하지만, 구직자의 가장 큰 관심사는 ‘상대적인’ 정보이다. 절대적으로 높아 보이는 기준일지라도 다른 기업들과 비교했을 때 상대적으로 낮은 수치라면 해당 기업을 선택할 이유는 없을 것이다. 대부분은 구인, 구직 사이트에서 채용정보와 기업정보를 확인한 후 다른 검색 사이트나 전자공시 사이트 등을 통해 기업의 상대적인 정보를 얻곤 한다. 글에 제시된 아이디어는 이러한 과정을 단순화시킬 수 있는 시스템이다. 현재 게시되어 있는 채용 및 기업정보와 더불어 각 기업 간 비교할 수 있는 시스템을 활용하여, 구직자는 원하는 정보를 찾는 시간을 줄이고 효율적으로 기업에 대해 이해하며 자신이 설정한 기준에 적합한 일자리에 지원할 수 있게 된다.

(2) 사회적 측면

채용정보 비교시스템을 구축할 때 ‘중소기업’을 대상으로 모델링을 한 이유는 중소

기업의 인력난을 심각한 고용문제로 파악했기 때문이다. 이러한 문제가 일어나는 이유는 다양하겠지만, 사람들의 선입견과 정보 부족이 큰 영향을 끼친다고 생각한다. 한때 중소기업의 좋지 않은 처우를 보여주는 유튜브의 ‘쫄쫄소’라는 드라마가 크게 인기 있었던 것은 많은 사람이 그 열악한 환경에 공감하고 있음을 보여주는 좋은 사례이다. 실제로 소속 근로자에 대한 대우가 좋지 않은 일부 중소기업이 있겠지만, 그렇지 않은 훌륭한 중소기업 역시 존재한다. 하지만 시간이 지날수록 자연스레 사람들의 뇌리에 중소기업에 대한 부정적인 인식이 쌓이고 있다.

이를 해결할 수 있는 것은 중소기업에 대한 많은 정보를 보여주는 것이다. 임금, 복지 등 다른 기업들과 비교할 수 있는 정보를 제공하며 중소기업에 대해 충분히 이해시킬 필요가 있다. 구직자들이 선입견을 품고 중소기업을 파악하는 것이 아니라, 제공된 여러 기준을 활용해 종합적으로 기업을 판단하기 위한 도구를 제시하고 싶었고 이것이 바로 본 아이디어의 핵심이자 목적이다.

고용노동부의 청사진은 ‘국민 누구나 원하는 일자리에서 마음껏 역량을 발휘하는 나라’이다. 구직자들이 원하는 일자리를 찾기 위해서는 많은 선택지가 있어야 하고, 선택지를 쉽게 비교할 수 있어야 한다. 본 시스템을 통해, 자신이 진정으로 원하는 일자리에서 행복하게 일할 수 있는 세상에 한 발자국 더 다가가길 고대한다.

4. 참고문헌

1. 고용노동부, 2021년 한권으로 통하는 고용노동정책, 2021.02
2. 박지훈. "중소기업 근로자 인력 유입방안에 대한 고찰." 중소기업과 법 1.2 (2010): 141-174. 중소기업 근로자의 고용안정을 중심으로.
3. 이준우, 박종미, and 백소영. "지체장애인의 구직 및 직업생활 경험에 관한 사례연구: 장애인 당사자와 기업 인사담당자를 중심으로." 한국사회복지교육 39.- (2017): 59-91.
4. 고용노동통계 포털 DB (공공데이터) [임금근로시간] → [산업·규모별 임금 및 근로시간] → [2020년 이후] 22년 4월 데이터 사용
<http://laborstat.moel.go.kr/hmp/tblInfo/TblInfoList.do?menuId=0010001100101102&leftMenuId=0010001100101&bbsId>
5. 공공데이터포털(<https://www.data.go.kr/index.do>)
-> 출처: 포털 검색창에 “산업중분류별규모별사고사망자수” 입력. 제일 최신 버전.
6. 2022 고용노동부 강소기업 명단
<https://www.work.go.kr/jobyoung/smallGiants/smallGiantsBasicMain.do#tmp>
7. 허정원, '취업 힘들어도 안가요'...중소기업 구인난 대기업의 7.8배, 고용노동부, 중앙일보, 2020.01.22
<https://www.joongang.co.kr/article/23688173#home>
8. 조영미, '구직자 강점 AI로 분석... 맞춤형 직업 매칭 “취업 걱정 끝!”', 부산일보, 2022.04.10
<http://www.busan.com/view/busan/view.php?code=2022041018250337156>